GAATME: A Genetic Algorithm for Adversarial Translation Metrics Evaluation

Josef Jon, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{jon,bojar}@ufal.mff.cuni.cz

Abstract

Building on a recent method for decoding translation candidates from a Machine Translation (MT) model via a genetic algorithm, we modify it to generate adversarial translations to test and challenge MT evaluation metrics. The produced translations score very well in an arbitrary MT evaluation metric selected beforehand, despite containing serious, deliberately introduced errors. The method can be used to create adversarial test sets to analyze the biases and shortcomings of the metrics. We publish various such test sets for the Czech to English language pair, as well as the code to convert any parallel data into a similar adversarial test set.

1. Introduction

One of the crucial aspects of developing and deploying machine translation is automatic evaluation. The evaluation metrics introduced in recent years follow the trend of using pre-trained large language models as the core of a task-specific system. These novel metrics correlate better with human evaluation than the previous generation of metrics based on a rather shallow similarity of the proposed translation and human reference. However, many shortcomings, weaknesses and blind spots of these new metrics were already described in the literature, like insensitivity to errors in the translation of named entities, numbers and others (Hanna and Bojar, 2021; Amrhein and Sennrich, 2022).

In this paper, we modify a recently introduced genetic algorithm-based technique (Jon and Bojar, 2023) to automatically construct adversarial examples for specific metrics. Starting with an initial set of translation hypotheses generated by an MT model, we stochastically modify and combine them. The objective is to craft translations that excel in one specific metric but perform poorly across others. The main contribution of this paper is the release of metric-specific adversarial test sets and the accompanying code for creating new test sets, enabling researchers to probe various metrics' robustness, biases, and weaknesses. The code and the test sets can be found at: `https://github.com/cepin19/GAATME`

2. Related work

Many new MT evaluation metrics were introduced recently (Zhang et al., 2020; Yuan et al., 2021; Thompson and Post, 2020; Sellam et al., 2020; Rei et al., 2020, 2021, 2022b; Lo, 2019; Wan et al., 2021, 2022; Freitag et al., 2022; Rei et al., 2022a; Kocmi and Federmann, 2023; Guerreiro et al., 2023). They are based on representing the source, MT, and reference sentences in a (sometimes shared) high-dimensional space, computing the similarity between the representations, and (in most cases) predicting human evaluation scores. This allows more flexibility than traditional metrics based on shallow text similarities (e.g. for lexical overlap metrics like BLEU, synonyms vs. completely unrelated mistranslations are indistinguishable, while neural metrics should account for this by scoring synonyms similarly). Overall, they correlate better with human evaluation (Freitag et al., 2022; Kocmi et al., 2021). The downside of this increased flexibility is that the models are prone to be insensitive to some kinds of errors, especially in rare words and named entities, since such expressions often have similar embeddings.

Existing literature extensively probes the behavior and weaknesses of these contemporary metrics. Moghe et al. (2023) show that neural metrics do not provide reliable results on the segment level. Amrhein and Sennrich (2022) try to find high-scoring incorrect translations, similar to our approach, to show that the analyzed metrics are not sensitive to errors in named entities and numbers. Alves et al. (2022) and Kanojia et al. (2021) further show that meaning-changing errors are hard to detect for QE. Rei et al. (2023); Leiter et al. (2022); Treviso et al. (2021); Guerreiro et al. (2023), and Fomicheva et al. (2021) explore the interpretability of the neural metrics.

Lastly, we directly build on Jon and Bojar (2023) and use a slight modification of this approach (using negative weights for "control" metrics) to build our adversarial test sets.

3. Method

We adapt the method presented by Jon and Bojar (2023). This approach is based on the genetic algorithm, which is described in the following paragraphs.

LREC-COLING 2024, pages 7562–7569

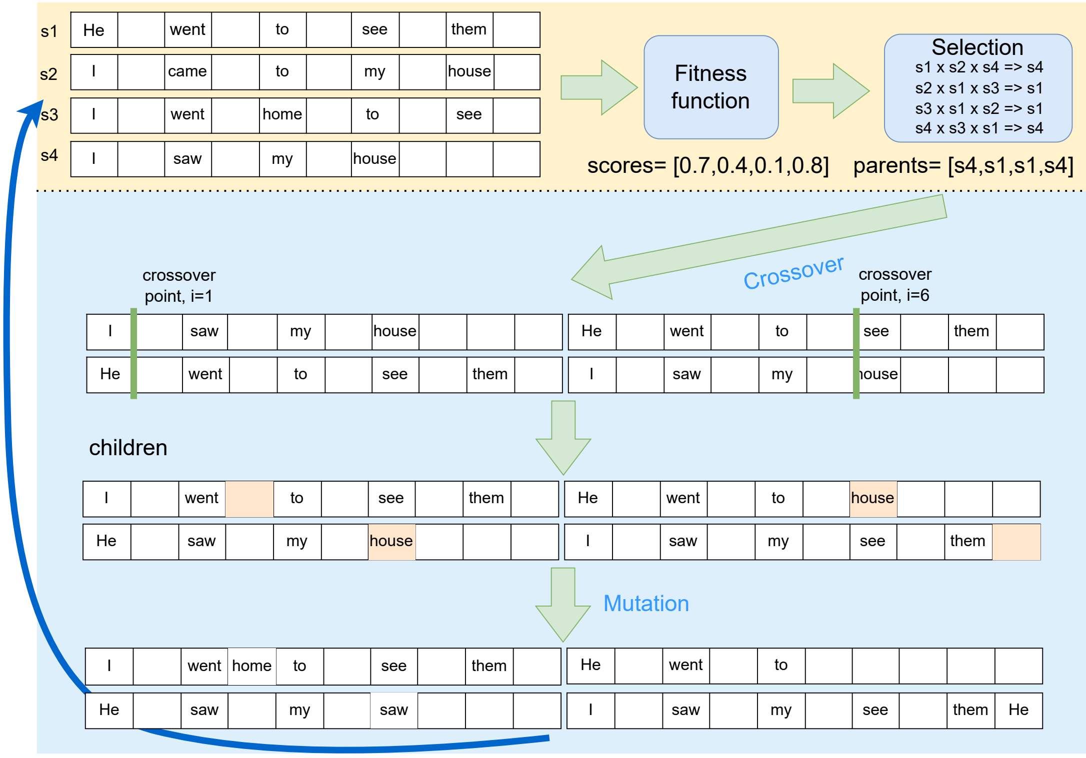

Figure 1: One iteration of the GA algorithm for a population of 4 individuals s1,. . . , s4. The steps with the yellow background are equivalent to simple reranking, the steps with the blue background introduce the operations of the genetic algorithm. Figure from Jon and Bojar (2023).

Genetic algorithm Our approach is the same as Jon and Bojar (2023) and we encourage the reader to find a more detailed description there. The whole process is illustrated in Figure 1.

First, a set of candidate sentences is produced by an NMT model, either by a beam search decoding or sampling for an increased diversity. These candidates are stochastically combined using a cross-over operation. This operation selects two individuals (i.e. translation hypotheses) from the population, splits them at a random token index swaps the split parts between the individuals. The resulting sequences are further modified using the mutation operation, which randomly removes, adds or replaces tokens in the candidate. The choices for new tokens to add or replace come from two sources: the complete wordlist in the target language and the set of words from a reference sentence.

Then, these candidate translations are scored using a fitness function, in our case a weighted sum of MT metrics' scores computed with regard to a known reference. This is a difference from Jon and Bojar (2023), who use MBR decoding with the translation candidates themselves as pseudoreferences. We use a positive weight for the metric we want to analyze (i.e., the one we want to find adversarial examples for) and a small negative weight for several other MT metrics. Our goal is to find sentences that perform exceptionally well according to the metric of interest but not as well according to other metrics. Finally, a new population of candidates is selected based on their scores via tournament selection.

In most of the experiments by Jon and Bojar (2023), the authors use MBR decoding to obtain the scores, in order to improve the quality of the translation. In our experiments, we do not keep the reference secret, since we are looking to obtain adversarial examples for the MT metrics instead. Jon and Bojar (2023) also ran a similarly designed small-scale experiment but they only used the analyzed metric for the fitness function, without the negatively weighted "control" metrics. They searched for the "suspicious" final translations in the outputs after running the whole GA algorithm. We are encouraging GA to directly prefer the suspicious translation candidates, making our approach proactive in seeking out translations that may reveal weaknesses in MT metrics.

4. Experiments

4.1. Data, tools and model

The NMT model was trained on CzEng 2.0 (Bojar et al., 2016; Kocmi et al., 2020) We obtained the English wordlist from `https://github.com/dwyl/english-words`. We tokenize the text using SentencePiece (Kudo and Richardson, 2018) and FactoredSegmenter[1] for the training. For the tokenization in the GA process, we use SacreMoses.[2] We used the `wmt22` (Kocmi et al., 2022) test set in Czech to English direction to create the adversarial translations. To produce the initial translations, we use the same model as Jon and Bojar (2023), i.e. `transformer-big` (Vaswani et al., 2017) using MarianNMT (Junczys-Dowmunt et al., 2018) with default hyperparameters.

4.2. GA parameters

The initial population of translation candidates is created by the NMT model described in Section 4.1. We concatenate n-best list obtained by beam search with beam size 20 and 20 sampled translations. We sample uniformly from the whole output distribution, as default in MarianNMT. We copy these 40 candidates 50 times to reach a population size of 2000. Empty token positions are added before and after each token in each candidate to support the addition of new words at these positions by mutating them to non-empty tokens. Finally, all the candidates are padded with empty token positions to the length of the longest candidate multiplied by 1.1.

The candidates are combined at a crossover rate of $c = 0.1$. The mutation rate for modifying non-empty genes (tokens) to other non-empty genes is $m = \frac{1}{l}$, with l representing chromosome length (i.e. number of positions in the translation candidate, including the empty token placeholders). Mutation rates between empty and non-empty genes (word addition/deletion) are $\frac{m}{10}$. Parents for the next generation are chosen via a tournament selection with $n = 3$. The GA stops after 150 generations. We combine the studied metric (the one we aim to find adversarial examples for) with other metrics in the fitness function by a weighted sum. We set the weight of the studied metric to 1.0 and we explore the following weights for all the other metrics: $0, -0.001, -0.01, -0.02, -0.03, -0.05, -0.1$. We note that these settings are arbitrary, based on some previous experience. A search for better parameters could bring further improvements, but running the whole process is computationally costly. This is mainly due to a need for running a large number of evaluations by the MT metrics, many of which are deep learning-based and resource-intensive.

4.3. Metrics

We assess the translations using the metrics: BLEU (Papineni et al., 2002), ChrF (Popović, 2015), BLEURT-20 (Sellam et al., 2020), wmt20-comet-da (CMT20 in the tables), wmt22-comet-da (CMT22), wmt22-cometkiwi-da (CMT22-QE) (Rei et al., 2020, 2022a,c) and UniTE-MUP (Wan et al., 2022).

For both BLEU and ChrF metrics, SacreBLEU (Post, 2018) is used. Specifically, ChrF operates with a $\beta = 2$ setting, labeled ChrF2.

BLEURT is not used as the negative metric in any of the experiments, due to its $\tilde{5}$x computational requirements compared to COMET. We only analyze it as the studied metric, with other metrics as the negative ones. We also do not use wmt20-comet-da as part of the negatively weighted metrics, because we previously found that it does not correlate well with human quality assessment under these circumstances.

4.4. Results

The results from our various experimental runs are summarized in Table 1. The first column specifies the metric we are creating adversarial examples for. The second column details the negative metric weights. These "control" metrics guide the GA to produce translations with errors. The following columns provide system-level scores of the adversarial translations produced. A translation is identified as adversarial if its post-GA score in the examined metric rises, while the translation manifests serious translation mistakes introduced by the GA process, as we manuall annotated, see below. The first row shows the results of the baseline MT model that was used to create the initial population of translation candidates for the GA.

We can infer some notions about the robustness of the particular metrics based on this table. By comparing the targeted metric's score with other (control) metrics' scores, we can get a gist of its resilience against adversarial inputs. If a metric can be tricked using our method, its post-GA score should remain high, whereas scores from other metrics should decrease significantly. For instance, when optimized for BLEU or CMT22-QE, we observe a decline in most other metrics compared to their baseline, even without negative weights in the fitness function. In other words, BLEU and CMT22-QE are very susceptible to overfitting towards them. Conversely, optimization for UniTE or CMT22 enhances scores in many other metrics, indicating the robustness of UniTE and CMT22. This kind of analysis assumes there are no spurious correlations or shared blind spots between the metrics – this assumption is however certainly violated in practice, since the neural metrics share large parts of the architecture and training data.

To address this, we manually examined a selection of translations to determine the true ratio of adversarial samples. We evaluated 50 samples from each metric with negative weights of 0 and -0.1, labeling them adversarial if they presented significant translation errors (consistent samples were

[1] https://github.com/microsoft/factored-segmenter

[2] https://github.com/alvations/sacremoses

Adv metric	$\mathbf{w_{neg}}$	ChrF	BLEU	CMT20	CMT22	CMT22-QE	BLEURT	UniTE	% better	% adv
Baseline MT		64.1	39.9	0.434	0.794	0.751	0.671	0.123		
	0	**84.8**	58.2	-0.220	0.634	0.508	0.543	-0.433	100	78 (78)
	0.001	**85.2**	58.6	-0.240	0.635	0.511	0.539	-0.427		
	0.01	**85.0**	56.4	-0.433	0.589	0.472	0.495	-0.668		
chrF	0.02	**84.8**	55.1	-0.640	0.544	0.435	0.444	-0.829		
	0.03	**84.4**	52.1	-0.815	0.517	0.415	0.408	-0.991		
	0.05	**82.6**	45.0	-1.078	0.460	0.381	0.363	-1.193		
	0.1	**79.9**	33.6	-1.257	0.406	0.341	0.353	-1.374	96	96 (100)
	0	74.8	**63.4**	0.122	0.730	0.607	0.588	-0.157	100	70 (70)
	0.001	71.4	**63.0**	-0.535	0.570	0.485	0.445	-0.789		
	0.01	69.2	**62.5**	-0.907	0.477	0.428	0.389	-1.099		
BLEU	0.02	68.7	**62.7**	-0.972	0.452	0.404	0.362	-1.149		
	0.03	67.5	**62.1**	-1.022	0.438	0.397	0.356	-1.187		
	0.05	65.7	**61.3**	-1.186	0.401	0.370	0.321	-1.321		
	0.1	61.8	**59.2**	-1.292	0.363	0.328	0.292	-1.427	98	98 (100)
	0	70.4	49.6	**0.803**	0.851	0.739	0.727	0.401	100	22 (22)
	0.001	69.9	49.0	**0.800**	0.850	0.733	0.719	0.357		
	0.01	70.7	49.7	**0.801**	0.849	0.723	0.701	0.299		
CMT20*	0.02	69.1	48.2	**0.799**	0.843	0.721	0.692	0.239		
	0.03	67.8	45.0	**0.794**	0.839	0.698	0.680	0.139		
	0.05	64.1	37.7	**0.772**	0.820	0.673	0.641	-0.016		
	0.1	55.5	24.0	**0.716**	0.779	0.599	0.561	-0.449	82	78 (96)
	0	69.4	48.4	0.667	**0.879**	0.742	0.710	0.311	100	26 (26)
	0.001	69.5	49.3	0.649	**0.879**	0.739	0.715	0.294		
	0.01	64.6	40.4	0.580	**0.876**	0.715	0.681	0.069		
CMT22	0.02	59.3	29.9	0.471	**0.867**	0.658	0.621	-0.225		
	0.03	53.3	22.6	0.259	**0.854**	0.612	0.569	-0.532		
	0.05	45.8	12.3	-0.071	**0.828**	0.535	0.490	-0.908		
	0.1	35.5	2.3	-0.518	**0.788**	0.449	0.401	-1.194	38	38 (100)
	0	61.2	35.3	0.400	0.809	**0.824**	0.674	0.080	100	20 (20)
	0.001	61.5	35.2	0.404	0.807	**0.822**	0.667	0.098		
	0.01	56.9	28.2	0.220	0.775	**0.819**	0.629	-0.157		
CMT22-QE	0.02	50.5	17.8	-0.217	0.711	**0.810**	0.551	-0.526		
	0.03	46.9	12.4	-0.461	0.670	**0.801**	0.510	-0.754		
	0.05	40.6	6.9	-0.770	0.601	**0.783**	0.449	-1.038		
	0.1	32.7	2.3	-1.079	0.515	**0.752**	0.402	-1.236	30	30 (100)
	0	65.1	40.8	0.048	0.721	0.611	**0.822**	-0.241	100	78 (78)
	0.001	65.0	40.4	0.076	0.732	0.638	**0.819**	-0.211		
	0.01	61.1	35.1	-0.374	0.633	0.515	**0.819**	-0.612		
BLEURT*	0.02	58.9	28.3	-0.665	0.567	0.459	**0.817**	-0.855		
	0.03	53.2	20.3	-0.814	0.531	0.429	**0.809**	-0.985		
	0.05	49.0	15.8	-1.008	0.480	0.401	**0.806**	-1.143		
	0.1	36.4	4.9	-1.275	0.406	0.348	**0.833**	-1.359	76	76 (100)
	0	68.4	45.4	0.591	0.817	0.726	0.707	**0.628**	100	22 (22)
	0.001	68.3	44.8	0.555	0.808	0.719	0.707	**0.622**		
	0.01	67.5	45.1	0.588	0.810	0.717	0.706	**0.643**		
UniTE	0.02	67.8	45.1	0.609	0.821	0.723	0.715	**0.636**		
	0.03	67.3	43.5	0.548	0.808	0.723	0.702	**0.622**		
	0.05	66.3	41.5	0.544	0.804	0.705	0.692	**0.615**		
	0.1	62.7	33.9	0.471	0.783	0.687	0.665	**0.610**	100	44 (44)

Table 1: Average scores of the generated test sets. Metrics marked with * were not used in the negative component of the fitness function for analyzing the other metrics. Scores in analyzed metric (the one we are searching adversarial examples for) are bold. ChrF and BLEU scores are multiplied by 100, in the algorithm they are in the 0-1 range. Higher is better for all the metrics.

used across all settings). Significant errors are defined as omissions, misinterpretations, additions not related to the source, redundant repetitions, or severe word-order errors. Our analysis is summarized in the last two columns. The "% better" column displays cases where the post-GA metric score surpasses its pre-GA value. The final column highlights instances that meet the previous criterion but also introduce a significant error via the GA. The presence of these errors was manually assessed. The numbers in parentheses show the total percentages of examples that contain newly introduced errors, regardless of whether the GA has improved the score or not.

4.5. Examples

Examples from our final test sets are presented in Table 2. Each row of the final test set displays the name of the analyzed metric, the source sentence (which is omitted in the table for conciseness), the machine translation (the first translation from the n-best list used as the initial population), the best translation post-GA, as well as the pre-GA and post-GA scores for the analyzed metric. This table offers insights into common errors associated with specific metrics. For instance, BLEURT appears to favor unfamiliar words or terms from other languages. These words were part of the noise in the English wordlist. We were unaware of their presence in the

Metric	MT	post-GA	Ref	MT score	GA score
CMT22-QE	In the NHL, "France" caught 36 **games**, its **save** rate at 92.3%.	In **yn**, "Frederic" clocked up 36 **ordain**, with **touchdown** rate at 92.3 **Basilica**	He has played 36 games in the NHL, where his save percentage is 92.3%.	0.6407	0.7679
	The 31-year-old full-back will be on the **scoresheet** and could soon be in goal.	The fullback will **toilette on rotation** and could get into goal soon **fungo**	The thirty-one-year-old Pilsen native will be on the bench and could soon be in goal.	0.6901	0.7242
CMT22	The highest ranked in the affair is **Berbr**, who no longer features in any of the football functions.	The highest profile in **glave** affair is **Piute Denten** who no longer longer figures **stanno** any football functions	The most senior figure in the affair is Berbr, who is no longer involved in any football function.	0.7947	0.8104
	Prince **William, Duke of Cambridge,** is wearing the same as Princes **George** and Louis shorts and a collared T-shirt.	Prince **Pippo, Duke of Goldwyn**, dressed the same as Princes **Alexander** and Louis in shorts and a T-shirt	Prince William, Duke of Cambridge, and Princes George and Louis are wearing shorts and a polo shirt.	0.7675	0.8402
CHRF	Interior has got respirators **significantly** cheaper than the Department of Health	Interior got respirators **mushy Asch** cheaper than the Ministry **oie** Ministry of Health **natl . fur . LADT Goethe**	The Ministry of the Interior got respirators much cheaper than the Ministry of Health	0.5342	0.8168
	PVO: medium-cold war, outdated; short range - good, modern, relatively good number.	**unpaint:** medium-cold war, obsolete; short range - good, modern, relatively **Orth . enterable** number**. fugitively favorer POS SMDF R.A.A.F. pm . SM**	SHORAD: medium - cold war, obsolete; short range - good, modern, relatively favorable number.	67.9	82.0
BLEU	**The picture**, which will serve as a Christmas card, was also posted by heir to the throne Prince Charles and wife Camilla.	**knotty-leaved**, which will be **fat-shunning weeny-bopper** Christmas card, was also posted by the heir to the throne, Prince Charles, **dichlorodiphenyltrichloroethane duodenocholecystostomy cock-a-doodle-doos** his wife, Camilla.	The image, which will be used for the Christmas card, was also posted by the heir to the throne, Prince Charles, and his wife, Camilla.	34.3	68.6
BLEURT	**By the time I** got off **my seat, it was** gone.	**Idun epicanthi** got **achenium tundun terebinthial** off **Ladakhi Morgenthaler** gone.. **ecliptically scholium mesonasal**	By the time I got off the deer-stand, he was gone.	0.3965	0.8183
	His return to goal in the NHL **eventually extended to more than two** months.	**succinimid Badajoz hootchie-kootchie cheongsam** NHL **tao-tai meromyarian Abyla Nadean vainer tenson** months	In the end, his time away from the NHL was extended by more than two months.	0.5695	0.8510

Table 2: Examples from the adversarial test set. Superfluous words in the post-GA translation and words from before GA that are missing post-GA are in bold.

wordlist before running the experiments.

5. Conclusion

We have presented a method to automatically generate adversarial test sets for arbitrary evaluation metrics. Our results demonstrate that this method is capable of producing translations that, while scoring higher than initial (correct or at least reasonable) MT outputs, contain serious translation errors. We have found that robustness against this method varies between metrics, with `wmt22-comet-da` and `UniTE` being particularly robust, while BLEURT (alongside BLEU and CHRF) can be surprisingly easy to deceive. We publish the code and the created test sets to allow further use of this method.

6. Acknowledgements

This work was partially supported by GAČR EXPRO grant NEUREM3 (19-26934X), by the Grant Agency of Charles University in Prague (GAUK 244523) and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

7. Bibliographical References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust mt evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation*, pages 469–478, Abu Dhabi. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. Czeng 1.6:

Enlarged czech-english parallel corpus with processing tools dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Springer International Publishing.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu â€“ neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Josef Jon and Ondřej Bojar. 2023. Breeding machine translations: Evolutionary approach to survive and thrive in the world of automated evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2191–2212, Toronto, Canada. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

Papers), pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022c. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Marcos Treviso, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the explainable quality estimation shared task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 133–145, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. RoBLEURT submission for WMT2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1053–1058, Online. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F. Wong, and Lidia S. Chao. 2022. UniTE: Unified Translation Evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.